Botnet C&C Detection Based on Machine Learning

Junzhi YAN
China Mobile

Emerging Botnet Attacks

MIRAI
2016.10

- Large scale of internet service disruption
- Cameras and DVR players are affected
- Simultaneous DDoS attack rate up to 1.5T
- Nearly 100 weak passwords are utilized

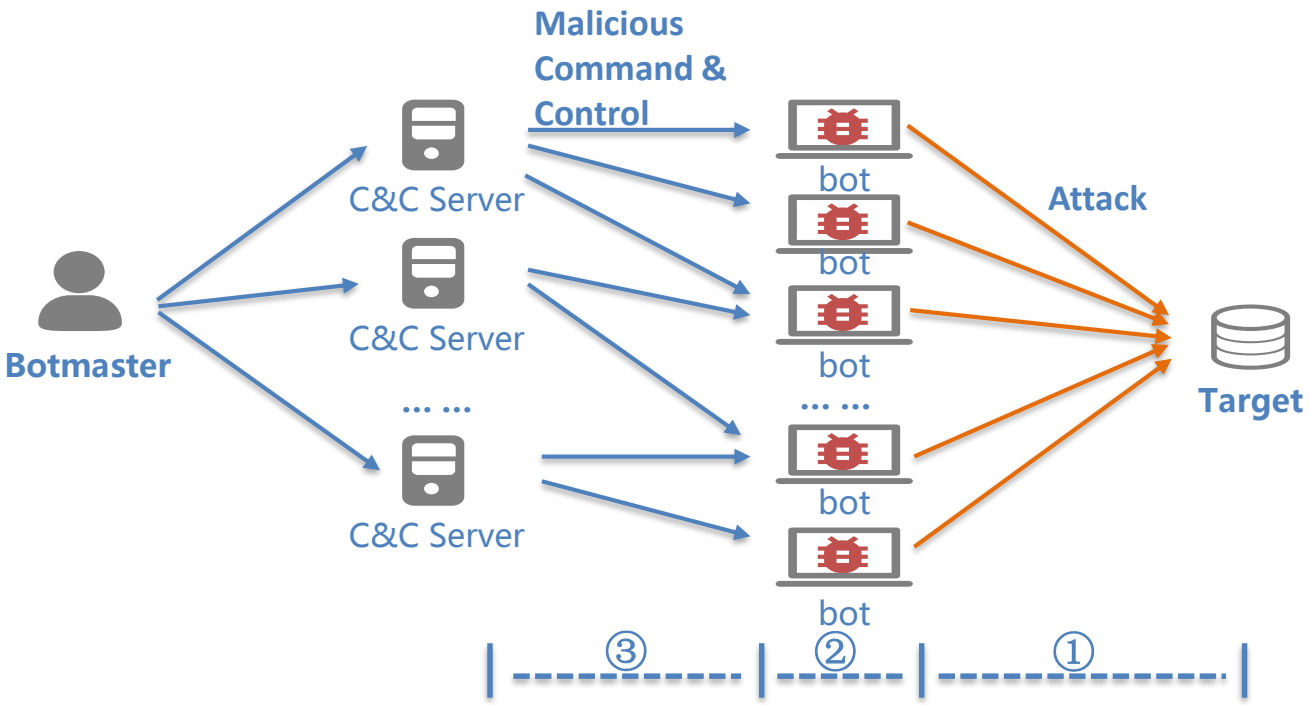
Reaper
2017.10

- Millions of companies are affected
- Vulnerabilities of IoT devices from different device makers are utilized

IoT Mining
2018

- Thousands of smart phones and TVs became Miners
- CPU and memory resources of the affected devices will be exhausted
- ADB.Miner: Android 5555 port

Botnet and Botnet Detection



① **Anomaly based detection:** high network latency, high volumes of traffic, traffic on unusual ports; un-efficient if the botnet has not been used for attacks

② **Signature based detection:** to find the signs of intrusion, using rules or signatures to find suspicious traffic; useful for detection of known botnet but unknown attacks

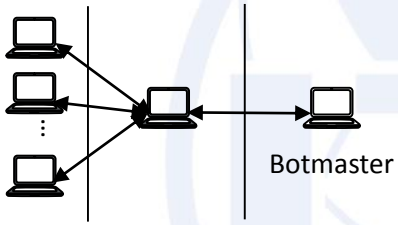
③ **DNS based detection**: to find unusual domain names, and detect DNS traffic anomalies

Structures of Botnet (1)

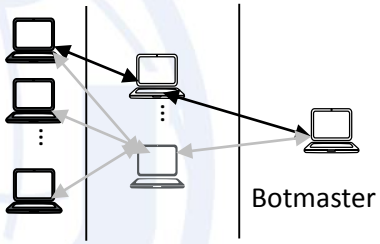
1. Centralized Structure

Client-Server mode is utilized, the bot should connect the C&C server to get the commands. The structure could be divided into static and dynamic one.

- **Static:** the C&C server's address is not changed and is programed into the malware
- **Dynamic:** the address of the C&C server is derived dynamically



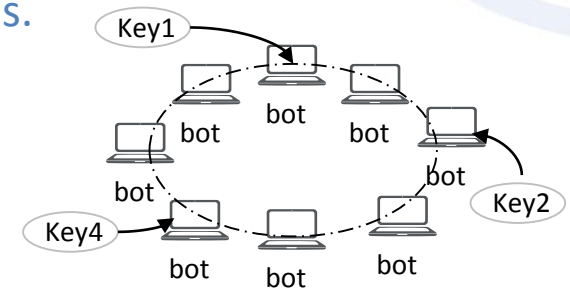
Static Structure

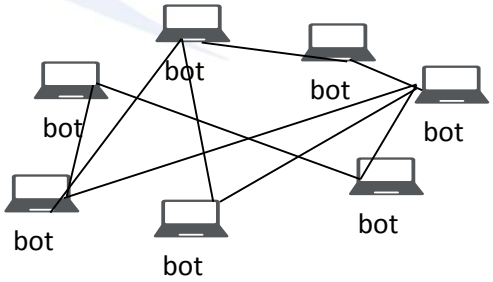


Dynamic Structure

2. Peer to Peer Structure

Each bot is a server and a client. The single point of failure could be avoided. Every bot has to find a peer, by using distributed hash table, random method, or some other methods.





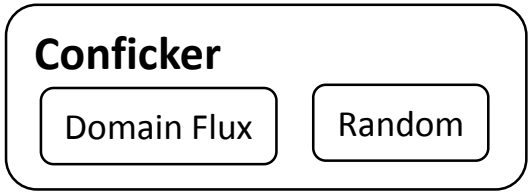
Structures of Botnet (2)

3. Combinatorial Structure

- C&C servers are P2P connected, and bot connects C&C server using C/S mode.
- Botnet are divided into sub-botnets, while these sub-networks are P2P connected.

4. Composite Structure

More than one structure are utilized. The availability of the C&C channel could be enhanced. Conficker, a worm in 2008, used domain flux (centralized structure) and random (P2P structure) together. Once domain flux does not work, random method could be used to find the peers.



Generation of C&C Server Domain Names

1. Single address

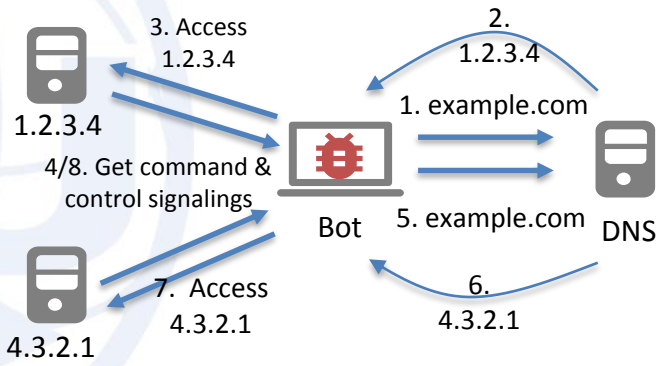
- The IP address or domain name of the C&C server is programed into the malware
- The IP address and domain name do not change, and the domain name is always long
- Could be detected easily

"network.bigbotpein.com"

Hostip=XXX.XXX.XXX.XXX

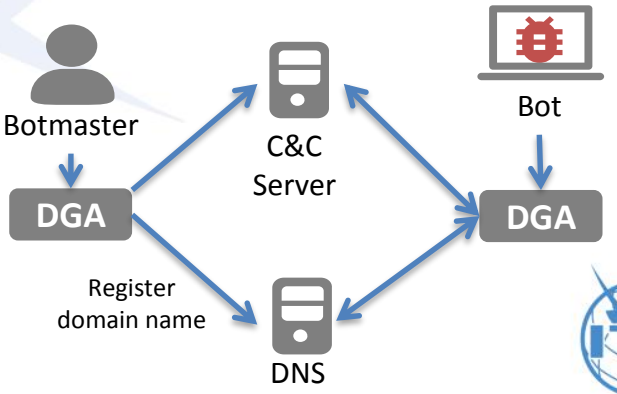
2. Fast flux

- Numerous C&C domain names and numerous IP addresses are used
- The IP address of the domain name changed frequently



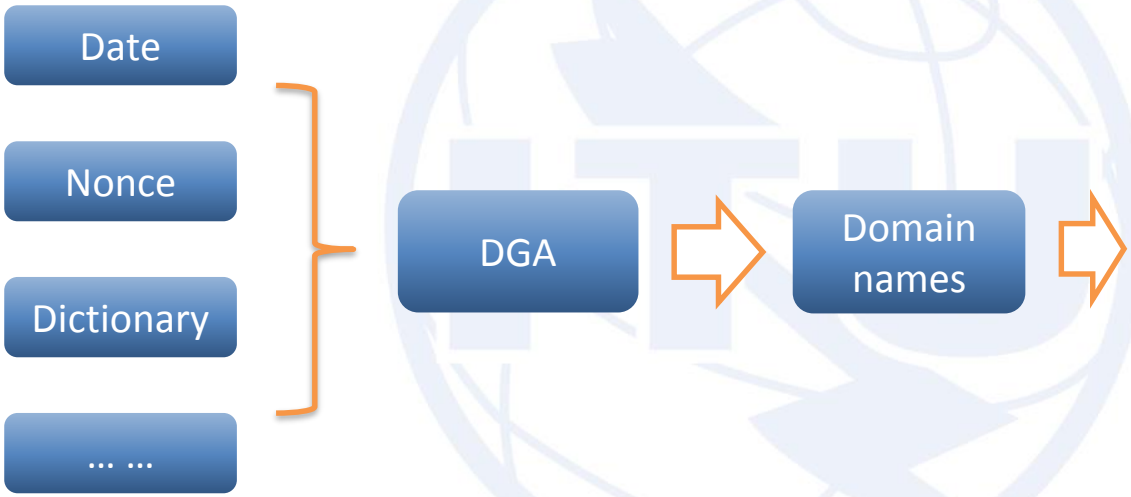
3. DGA

- No domain names in malware
- Candidate domain names are generated randomly, using coincident seeds
- Register the coincident domain name in DNS



Features of DGA Domain Names

Domain generation algorithms (DGA) are algorithms seen in various families of malware that are used to periodically generate a large number of domain names that can be used as rendezvous points with their command and control servers. ----Wikipedia



For Example:

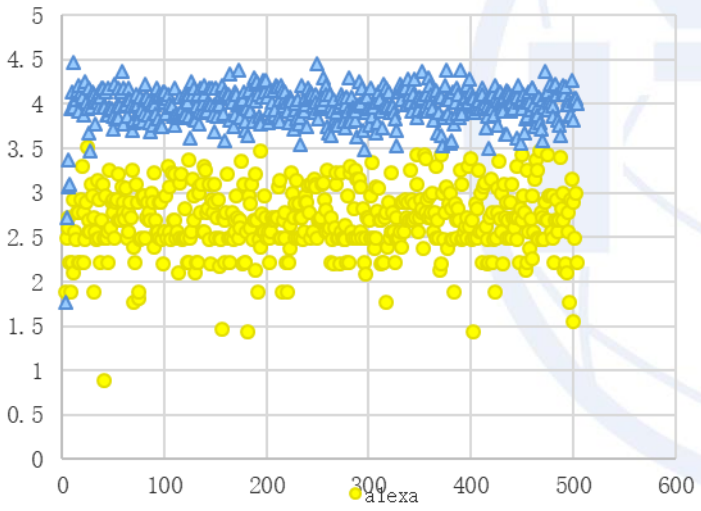
mootnxmn.org, ryrtthr.com
kbmeeou.biz, cuvcdergq.com
msnpuukp.org, cuxsjfxvmp.org
ctuyfbgirdq.com, eilcjtldgkj.org
mefklb.org, obdsccjdg.com
xocmyxvojfmvdueanjzxxwtcpzreq.org
xotwchepuwtyplbzaiqgofx.net
xilvgtxovqczhcqlzpnsgzpx.ru
ekvtestnessbiophysicalohax.com
skgxranansnan.com
… …

The DGA domain names are more random than normal domain names. Some features could be used to detect DGA domain names such as, domain name entropy, string length, vowel to consonant ratio, etc.

C&C Domain Name Detection

Entropy

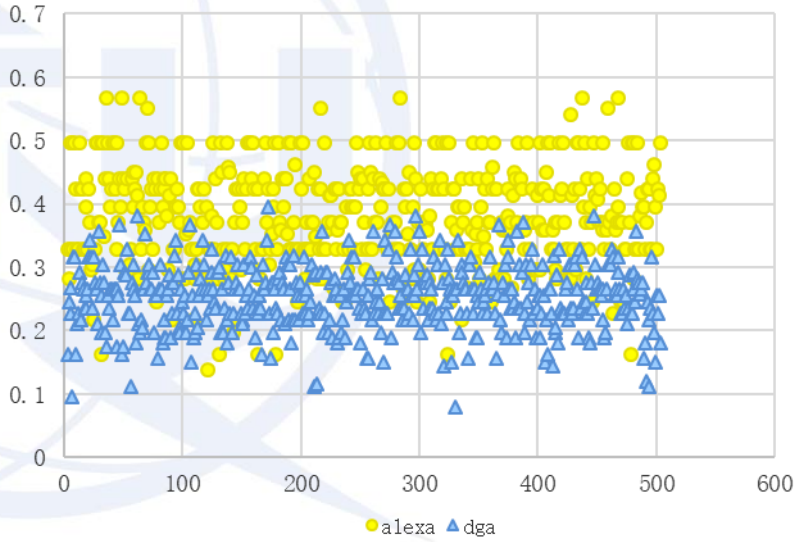
- Alexa: google-1.91
- DGA: 1vynq17gk6yo14ndoft2xc—4.07



- DGA domain names are more random, the entropy is larger than normal domain names

Consecutive Consonant

- Alexa: linkedin-1/8
- DGA: 1rudbfzc2z91r1ckvw5ge85u5d-2/36

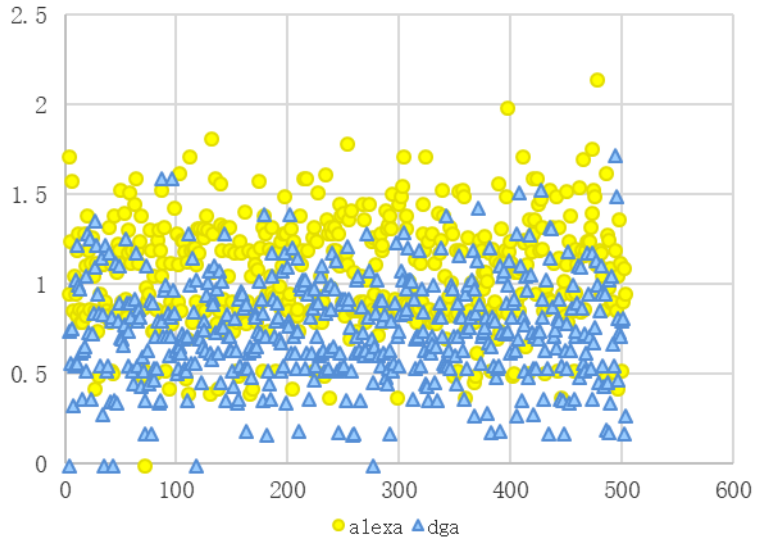


- Consecutive consonant to the length of domain names
- The ratio of DGA domain names is much more dispersed

C&C Domain Name Detection

Vowel ratio

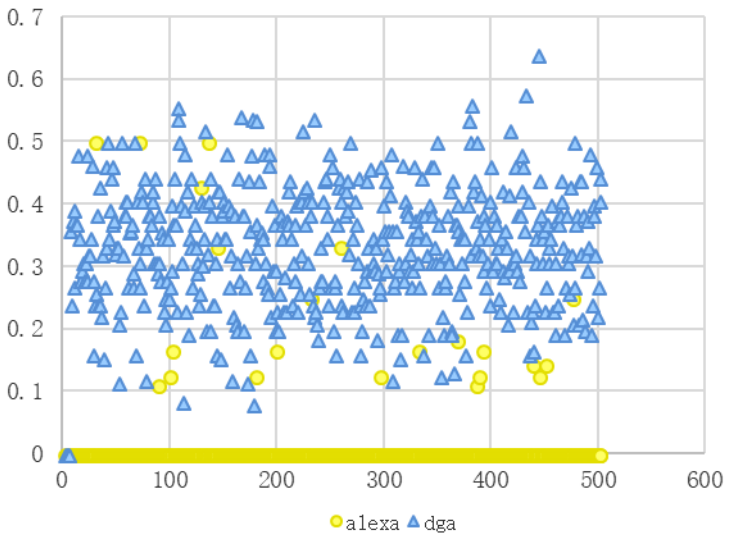
- Alexa: youtube-4/7
- DGA: 1jp6emzv47wflsrq0i1nt839-2/34



- The ratio of vowels to the length of domain names
- The ratio of vowels in DGA domain names is lower, and there is less repetition

Numeral ratio

- Alexa: Microsoft-—0
- DGA: kb9r036vvjfeq3576t1do282k—12/25

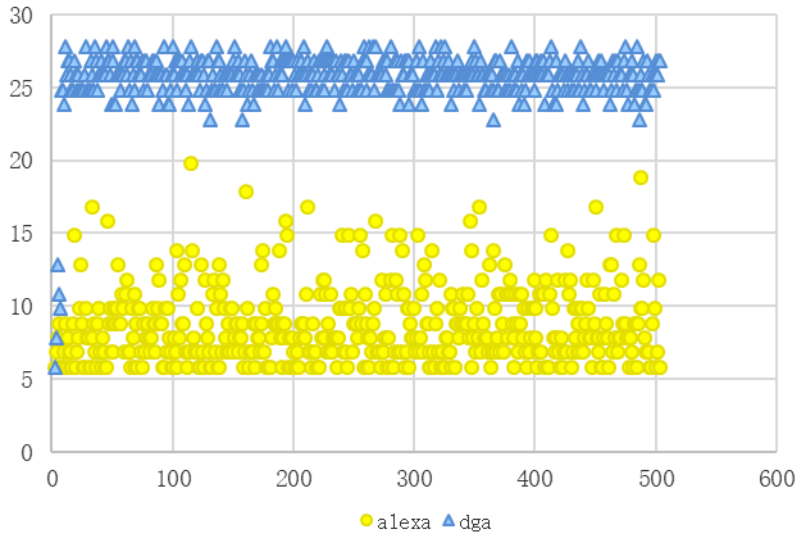


- The ratio of numerals
- The ratio in DGA domain names is much higher

C&C Domain Name Detection

Domain Name Length

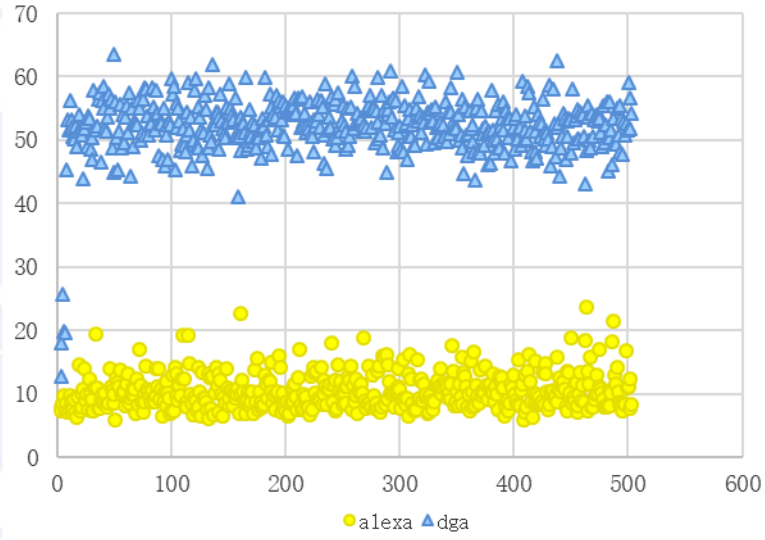
- Alexa: wikipedia-—9
- DGA: cfxajb17j9w7o5zh5l1c6hz5k—25



- The DGA domain name is longer

N-gram

- Alexa: facebook—11
- DGA: 5ms9mz1dpy3onvej4adswnw2v—53



- Trigram(N=3): probability of a word, conditioned on previous N-1 words

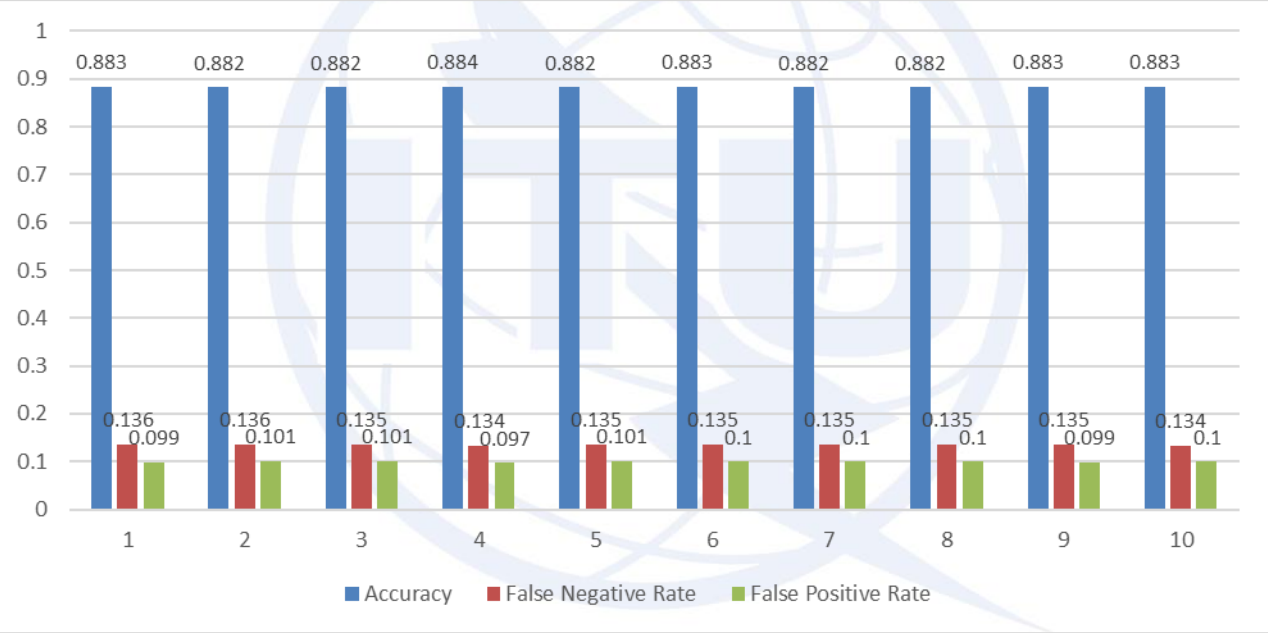
C&C Domain Name Detection

Positive Sample

- **1 Million** domain names from Alexa
- Duplicated domains are removed
- Domain names more than 6 words are selected
- The total number is **783,867**

Negative Sample

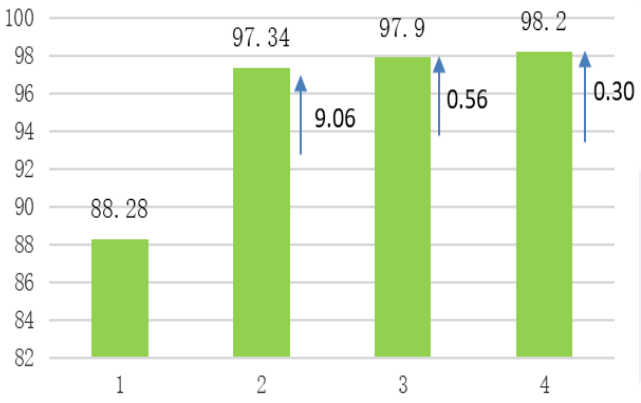
- **1,124,923** DGA domain names from netlab@360
- Duplicated domains are removed
- The total number is **1,063,695**



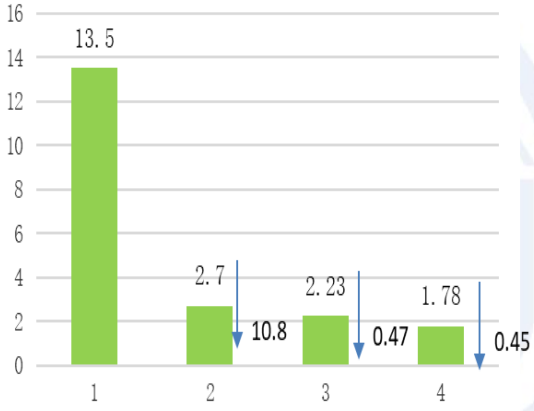
- **17 Features are fed into a machine learning model(random forest classier):** domain entropy, vowel entropy, bigram, trigram, etc.
- Average accuracy: 88.26%
- Average false negative rate: 13.5%
- Average false positive rate: 9.98%

C&C Domain Name Detection

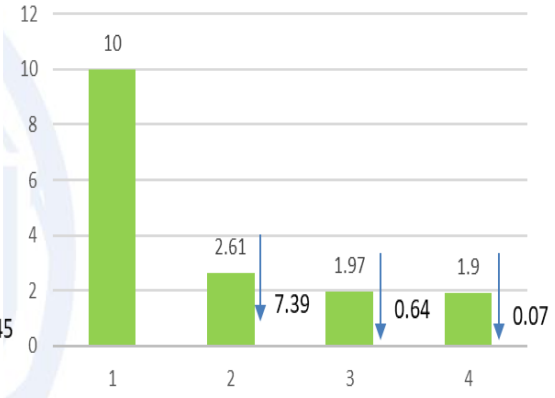
Accuracy



False Negative Rate



False Positive Rate



1: 17 Features

2: Optimization using N-gram normalization

3: Optimization using vector product

4: Optimization using consonant combination

Conclusion

- Machine learning is an efficient way to help detect the suspicious botnet C&C servers
- Combined with other methods, such DNS traffic anomalies, the efficiency could be enhanced
- Other methods are needed to confirm a suspicious server to be a malicious server

Thank you